2003:138 CIV

LULEÅ UNIVERSITY OF TECHNOLOGY

MASTER'S THESIS

Bimodal Voice Recognition Based Computer Input

WANG YU

MASTER OF SCIENCE PROGRAMME
M.Sc. Report in Industrial Ergonomics

Department of Human Work Sciences
Division of Industrial Ergonomics

2003:138 CIV • ISSN: 1402 - 1617 • ISRN: LTU - EX - - 03/138 - - SE

BIMODAL VOICE RECOGNITION BASED COMPUTER INPUT

Wang Yu

2003-03-14

Industrial Ergonomics

Department of Human Work Sciences

Luleå University of Technology

ACKNOWLEDGEMENTS

A number of people have been involved in my work and I wish to express my warmest gratitude to everyone who supported and help me in different ways.

First of all I am deeply grateful to my husband Jianlin Shi. Without his understanding, valuable advice and patience, my thesis could not have been attempted and completed.

I want to express my heartfelt thanks to my supervisor Emma-Christin Lönnroth for her persistent help, skilful and excellent guidance and positive encouragement. I really appreciated her enthusiasm and valuable suggestions during the whole period that I studied at the Division of Industrial Ergonomics of M.Sc. program.

Thanks to Professor Houshang Shahnavaz for introducing me into ergonomic field and valuable instructions.

My sincerely gratitude also goes to my friends Li Xin, Lui Hongyuan, Cui Jirang, Ma Haoxue and Wu peng for their kind helps and discussions on everything inside as well as outside the scientific world.

I would like to express my thanks to all of my colleagues in M.Sc. program for their help and friendship.

I wish to express my gratitude to my parent for their great support. Finally, I dedicate this thesis in the honour of my dear husband Jianlin.

Abstract

In the last few decades, the computer keyboards input device has received much attention in the past and is believed by many to be a prime factor in the etiology of upper extremity musculoskeletal disorders. And wide rang of voice input systems are proposed to allow persons to operate a computer without using a keyboard or mouse.

This thesis reviewed both of acoustic–only and bimodal voice recognition system and compared their recognition accuracy in simulated noisy environments. Then, the voice recognition technique is adopted in keyboard design to fulfil keyboard ergonomic demands. Finally, the value analysis was performed to evaluate the redesigned voice input keyboards.

The experiment results demonstrate, compared to conventional acoustic only based speech recognition, bimodal speech recognition scheme has a much improved recognition accuracy and using the visual features allows the development of a more practical and real-time recognition system. Through the redesigned voice input keyboard, computer users can get their hand free completely and partly at their own will, by which they are away from the upper extremity musculoskeletal disorders risk and vocal strain.

Keyword: upper extremity musculoskeletal disorder, keyboard, voice input, speech recognition.

Table of Contents

Nomenclature and Abbreviation

ASR	Auto Speech Recognition
HMM	Hidden Markov Model
EMG	Electromyography
NIOSH	National Institute for Occupational Safety and Health
BLS	the Bureau of Labour Statistics
WMSD	Work-related Musculoskeletal Disorders
WRUED	Work-related Upper Extremity Disorders
RSI	Repetitive Strain or Stress Injuries
RMI	Repetitive Motion Injures
CTD	Cumulative Trauma Disorders
CTS	Carpal Tunnel Syndrome
Pi	weight Factor of function i
Ki	weight Number of function i
RP_i	ranking value of proposal for function i
o	ordered sequence of the observation
o_t	observation vector at time t
q_t	state variable at time t
N	number of the states

M	number of the mixture components in a state
a_{ij}	transition probability from state i to state j
$b_j(o_t)$	observation probability o_t of finding in state j
λ	a set of probability parameters for a HMM
$\overline{\lambda}$	auxiliary variable corresponds to λ
π_i	initial state probability for state i
$\alpha_t(i)$	forward probability
$\beta_t(i)$	backward probability
$\delta_t(j)$	partial likelihood
$\psi_t(j)$	trace of the state sequence
$\xi_t(i, j)$	probability of being in state s_i at time t and state s_j at time $t+1$, given the o and λ
$\gamma_t(i)$	probability of being in state s_i at time t, given the o and λ
c_{jm}	the mixture coefficient for in state j
μ_{jm}	mean vector of the mixture component mth in state j
W_{jm}	covariance matrix of the mixture component mth in state j

List of Figures

List of Tables

1 Introduction

1.1 General Introduction

In the last few decades, computer usage has experienced exponential growth due to the broad usage of computers to maintain and access global databases and process the large volume of data associated with different kinds of industries and researches. Gerard et al. (1994) and William Lehr, (1998) show us the dramatic raise in computer usage focused on services and government agencies in US.

Unfortunately, the occurrence of musculoskeletal injuries has also risen greatly along with computer usage. According to the Bureau of Labour Statistics (BLS, 2000), musculoskeletal disorders are prevalent in the office due to computer work. In 1996 there were 73,796 nonfatal occupational injuries and illnesses involving days away from work due to the repetitive motion. Of these cases, 11,226 were directly attributed to repetitive typing or key-entry (BLS, 1996). And Gerr et al. (2002) indicates that over 50% of newly hired computer users reported musculoskeletal symptoms within the first year on a job. Symptoms include eyestrain, neck and shoulder pain, low back pain, elbow pain (tendonitis), forearm pain (muscles) and nerve entrapments. These cases are also known as work-related musculoskeletal disorders (WMSD), work-related upper extremity disorders (WRUED), repetitive strain or stress injuries (RSI) and repetitive motion injures (RMI).

Silverstein (1986) and Armstrong (1987) pointed that main risk factors related to these injuries were high force, repetition, awkward postures, and sharp contact pressures. These risk factors are all present while working on a computer using a keyboard. The increased repetitive motions and awkward postures attributed to the use of computer keyboards have resulted in a rise in cumulative trauma disorders (CTD) that are generally considered to be the most costly and severe disorders occurring in the office. Several studies have examined the relationship between keyboard usages, also commonly referred as VDT (Video Display Terminal) usage, and the development of CTDs. (Pascarelli and Kella, 1993; Smutz et al., 1994; Gerard et al., 1994; Tittiranonda et al., 1994; Fernström et al., 1994 hedge and Powers, 1995; Martin et al., 1996; Feuerstein et al. 1997).

As a result, more and more researchers proposed different ergonomic devices to displace the traditional keyboard to reduce the injury risk. Kinesis keyboard, Dvorak keyboard, Lexmark keyboard, and MS Natural keyboard are the typical ergonomic keyboards in current market, as it will be introduced in detail in next chapter.

Besides these ergonomic keyboards, there is also a more satisfying substitutable design---voice input. In no other area of assistive technology has recent development been as dramatic as in the area of speech recognition. Recent advances in computer technology have enabled users of speech recognition products to achieve desirable results which was previously impossible on any but the largest mainframe computers or workstations. As a result, large numbers of voice input systems are produced to computer uses. It is expected that speech will be poised to replace the physical manipulation as the dominant input modality. This shift will dramatically alter our input needs, and the way we interact with computers.

However, there are some limitations and shortcomings in current voice input systems. One of which is the recognition accuracy. Especially in a noisy environment the recognition accuracy will decrease greatly. It is because all these current recognizer are acoustic based speech recognition, which is sensitive to noise signal. Since voice input is a delicate procedure, a slight change in ambient noise can affect the recognition accuracy a lot.

Thus a new input design were proposed based on bimodal voice recognition, which adapts visual and acoustic information together to recognize. The primary advantage of this method is that the visual information is not affected by acoustic noise cross talk among speakers. The studies in human perception system have shown that visual information allows people to tolerate an extra 4 dB of noise in the acoustic signal (J.Movellan, 1995). Secondly, visual information may lead speaker independent recognition to a high accuracy. Another advantage is the complementary structure of phonemes and visemes, which are the smallest acoustically and visually distinguishing units of a given language respectively. The third advantage is that visual information helps to localize the speaker (audio source) and offer clear visual information that supplements the audio signal.

Therefore, in this thesis, a bimodal voice recognition based voice input is proposed and examined. The experiments results showed that this new method has an advantage over the current voice input method from an aspect of recognition accuracy. Based on our experiments, this thesis suggested two integrated ergonomic voice input devices which adopt the acoustic-only and bimodal speech recognition techniques.

1.2 Thesis Organization

The thesis is organized as follows. The first chapter briefly introduces the background of this thesis. Chapter 2 reviews and analyses the traditional keyboard and ergonomic designs including alternative keyboards and voice input. Chapter 3 describes the fundamentals of the speech recognition theory and the implementation of bimodal input. Chapter 4 illustrates the experiments aiming to compare our proposed method with the conventional method. Finally, chapter 5 gives the general discussion and draws conclusion in Chapter 6.

2 Keyboard and Ergonomic Input Design

2.1 Traditional QWERTY Keyboard

The traditional layout was first introduced by Christopher Latham Sholes and Glidden(1866) as the result of modification on typewriters and telegraph's keyboards. After more than one hundred years, it became the universal input keyboard layout even in the most advanced computers. Its layout consists of four parallel rows of keys that in sum comprise the 26 letters of the alphabet, 10 numeric keys, and several other specific symbol or function keys. All these are placed in four different sections:

- The central portion that consists of letter keys
- The small right hand section containing number keys
- The small set of function keys between the letters and numbers
- A row of function keys going across the top

It gets its name QWERTY Keyboard from the spelling of the first six letter keys on the second row of the keyboard.

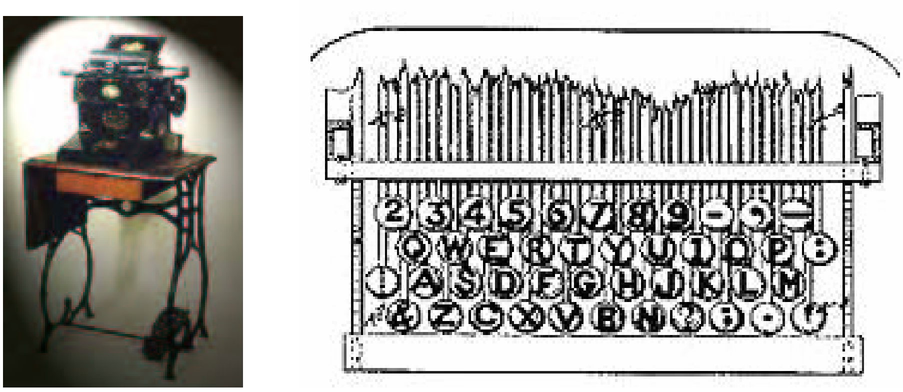

Figure 2.1 (a) Sholes & Glidden Typewriter of 1874; (b) 1878 Typewriter Patent Drawing, featuring the QWERTY Keyboard (http://www.library.wisc.edu/-

etext/WIReader/Images/WER0841.html)

2.1.1 Why Current Keyboard Need to Be Changed

Computer users frequently input data through the keyboard, and the conventional QWERTY keyboard has been used for more than 100 years without any modification. As a result, this input device has been the subject of much inquiry (Pascarelli and Kella, 1993; Smutz et al., 1994; Gerard et al., 1994; Tittiranonda et al., 1994; Fernström et al., 1994 hedge and Powers, 1995; Martin et al., 1996; Feuerstein et al. 1997).

Carter and Banister (1994) listed the possible caused and musculoskeletal injuries to VDT workers and these results are produced in four main categories: tendon disorders, nerve disorders, neurovascular disorders, and bone disorders. One possible factor contributing to CTD development that has been examined extensively is the keyboard. The possible causes related to keyboard issues are mainly awkward positions, static work, inactivity, overuse injury, stress on bone and connective tissue and pressure on blood vessels and nerves. And, keyboard positioning and layout are reported as important factor to force excessive ulnar abduction. (Bergqvist U.1995a b; Dennerlein JT, Yang MC.2001; Feuerstein M. et al., 1994; Gerard M.J. Gerard et al., 1994)

In order to determine the extent of the problem, the National Institute for Occupational Safety and Health (NIOSH) has performed several studies on keyboard users within the last decade (HETA89, 90). Table 2.1 shows the summarized findings of a Health Hazard Evaluation of cumulative trauma injuries among keyboard users that was conducted at Newsday, Inc. and Los Angeles Times. The results of this Health Hazard Evaluation by NIOSH revealed that 40% (in Newsday) and 41% (in Los Angeles Times) of the participating employees reported symptoms consistent with upper extremity cumulative trauma disorders.

Table 2.1 Musculoskeletal discomforts among keyboard users in Newsday and Los Angeles Times (http://www.aopd.com/vdt.html)

	Hand/wrist symptoms	Neck symptoms	Elbow/forearm symptoms	Shoulder symptoms
Newsday (89)	23%	17%	13%	11%
LA Times (90)	22%	26%	10%	17%

BLS reports yearly the number of repeated trauma illnesses increased rapidly in the past but peaked in 1994. The repeated occupational injuries and illnesses with time off from work due to trauma disorders are shown as following table.

There are more and more reports against the QWERTY keyboard. As a summary, alternative keyboards were purchased for the following reasons according to Kenneth Scott Wright and Dr. Anthony D. Andre's survey (1996) among keyboard users.

- Existing Injury/Pain (65%)
- Avoid Potential Injury (40%)
- Recommended/Provided (25%)
- Adjustable Design (23%)
- Disability Accommodation (17%)
- State-of-the-Art / Looked Cool(9%)

2.1.2 Main Reasons of Keyboard Injury:

Health hazard evaluations were performed at NIOSH in order to analyse the contribution of workplace ergonomic factors to musculoskeletal problems among computer users. The result data indicated that almost 40% of the variance in discomfort at carious body sites could be explained by ergonomic factors in the workplace. Among the ergonomic factors, issues about keyboard such as location, support and work surface are one of the primary areas lead to discomfort. Sauter et al. (1991) reported that discomfort increased with increase in keyboard height above elbow lever. Hunting, Laubli and Grandjean (Hunting, Laubli and Grandjean 1983) reported similar associations.

Pascarelli and Kella (1993) noted both internal and external ergonomic risk factors associated with keyboard usage that should be highly considered when analyze the relationship between VDT usage and the development of CTDs. And they summarised these factors into three main groups: postural risk factors, force risk factors and other risk factors, as shown in table 2.3.

Table 2.3 Internal and external ergonomic risk factors associated with keyboard use that should be considered when analyze the relationship between VDT usage and the development of CTDs (Pascarelli and Kella, 1993)

Category of CTD risk factors	Observation
A. Postural risk factors	a. Awkward wrist positions that individuals assume when typing b. Habit of extending and not using the non-dominant thumb when typing c. Leaning too far forward
B. Force risk factors	a. Striking the keys with excessive force
C. Other risk factors	a. The presence of pre-existing joint hyper mobility b. The tendency of individuals to prefer to use certain fingers excessively

2.1.2.1 Postural Factors

Typing requires a lot of side-to-side hand motion because the keys covered by each finger are arranged along a diagonal. The excessive ulnar abduction necessary to use the keyboard leads to awkward postures as the elbow is typically moved laterally. The mal-alignment between the fingers and the keys due to the anatomical shape of the hand and the length of the fingers are also typically addressed as problems with the current QWERTY design.

The first study of effect of keyboard on upper extremity muscle activity was conduced by Lundervold (1951). In his study, the increase in muscular activity with forearm pronation was observed. Later, Zipp et al. (1983) confirmed these results and added that ulnar deviation also contributed to the increased electromyography (EMG) activity.

Wrist posture is also an important factor related to musculoskeletal disorders. The usual monolithic keyboard requires that the hands be bent at an uncomfortable angle to the wrists. Hedge and Powers (1995) examined different wrist postures such as with or without arm/wrist support and using or not using a negative slope keyboard while working on a QWERTY keyboard. Their results showed that an average negative slope of 12° below the horizontal led to some positive affection because the slope keyboard significantly decreased the wrist extension.

The keyboard position is associated discomfort in all body regions except for the lower back and shoulders. If the keyboard is placed in a low place, it will produce a certain degree strain on the neck and upper back since the arms are suspended downward in this posture. Furthermore, as Carter and Bannister (1994) pointed, the extension at the wrists is also an awkward posture that is recognized as a risk factor for musculoskeletal disorders. From the view of ergonomic, the keyboard should be placed at a height that keep the forearms level and the wrists straight, in which operators avoid awkward posture. A general consensus is that the height of g-h keys should be the same height as the elbow. However, no one is certain it will not lead to any potential injury even in this position.

In 1997, University of California successfully measured fatigue by measuring twitch force of the muscle after electrical stimulation. The results illustrated that symptoms of subjective fatigue occurred within one hour of typing; and, subjective fatigue recovered over a time course of hours. Low frequency fatigue did not occur until the end of four hours of keyboard use. Although there was a trend toward increasing muscle fatigue with increasing angles of wrist extension the differences were not statistically significant (Chien-Yi Lu, 1997).

2.1.2.2 Force and Other Factors:

There are some physical factors such as finger travel, striking force, key motion and the repetitiveness of the task that related to potential injury. For an example, Typing on a standard keyboard requires a lot of hand motion up and down on. Since the little finger is shorter, it has to go further to reach its keys. The Office Ergonomics Research Committee (OERC) developed an approach for static key force measurement for consideration in future standards since there was no common standard methods of measuring static key force in the early 90s (Gerard et al, 1994). Based on this approach, Feuerstein (1994) and his colleagues have successfully measured both static key force (the force required to active a key switch) and keying force (the actual force being applied by a user) (Feuerstein M., and Hickey, P., 1994, Feuerstein, M., Hickey, P., and Lincoln, A., 1997). As suggested by the American National Standard for Human factors Engineering of Visual Display Terminal Workstations (ANSI/HFS 100-1988), the necessary key activation force in modern keyboard is normally below 0.5 N, with an upper limitation of 1.5 N. However, Feuerstein's results indicated that some users strike the keys two to five times harder than necessary to activate the key switches.

Another researcher Martin and his colleagues carried out similar study by examining the relationship between keyboard reaction force and electromyography (EMG). Similar results were drawn that keyboard users stroke keys with over 5 times the necessary force (Martin, B.J., Rempel, D.M., 1996, Martin BJ, Armstrong TJ, Foulke JA, Natarajan S, Klinenberg E., Serina E., Rempel D., 1996). Since type work is a highly repetitive task in the hand and wrist, the high level force, together with the over travel (the distance between the activation point and the key bottoming point), will easily lead to Repetitive Stress Injuries (RSI).

2.2 Alternative Keyboard Designs

Based on the injury analysis of keyboard input, good deals of efforts have been made in ergonomically designed keyboards in order to reduce finger travel and fatigue and to promote a more natural hand, wrist, and arm typing posture. A good many of more ergonomic keyboards with split and/or adjustable typing sections were proposed. (Smutz et al., 1994; Gerard et al., 1994; Thompson et al., 1990; Kreifeldt et al., 1989; Morita, 1989; Grandjean et al., 1985. The most notable alternatives were described by Dvorak (1943), Kroemer (1972) and Hobday (1988).) The main method of keyboard development was focused primarily on optimizing physical key characteristics, finger capability, and key arrangement. Some of these ergonomic keyboards also have alternative key layouts. All these alternative input devices provide the same or similar function of the traditional QWERTY keyboard. Studies have investigated the effects of some of these alternative keyboards on posture, comfort and performance. These studies reported that some alternative keyboards may reduce non-neutral wrist postures, may increase comfort for some users and may maintain close to or equivalent typing performance compared to conventional keyboards. These studies also showed that the effects of different alternative keyboard designs were not all alike. To date, the research is inconclusive in term of the effect of alternative keyboards on the incidence of upper extremely musculoskeletal disorder (UEMSD).

2.2.1 Split Keyboards

Split keyboard is the most common type of alternative keyboards. It makes up approximately 90% of the ergonomic alternative keyboards market. This kind of design aims to improve the ergonomic characteristics of the traditional QWERTY keyboard, while maintaining its basic shape and well-learned QWERTY key arrangement. This makes it easier for typists to switch to new keyboard designs, that assist in improving hand and arm postures, without learning a whole new typing skill. Split keyboard, as described by its name, is the keys are divided in the middle. The basic reason for splitting the keyboard is to eliminate ulnar wrist deviation, a suspect static position in the development of CTS.

Of these split keyboards two basic designs exist – fixed and adjustable. As the name implies fixed split keyboard allow for no adjustability. Adjustable splits allow the board to be adjusted to individual configurations. They can be complicated and may not be as rugged as the fixed; however, they do achieve their goal of alleviating the awkward postures.

2.2.1.1 Fixed-Split Keyboards

An early fixed-split keyboard was suggested by Kroemer in 1972. He used the increase in EMG activity to measure the forearm pronation necessary to place the hands flatly on the keyboard. Considering the excessive ulnar deviation as part of his justification, he suggested a new split key layout design to alleviate the postural stresses of the conventional keyboard layout.

One example of the fixed slit keyboard is Vertical Keyboards. It takes the standard keyboard's key sections and places them upright. This "hand-shake" position is considered the neutral posture for the forearms and hands. There are also some of the adjustable-split keyboards that can also assume vertical positions.

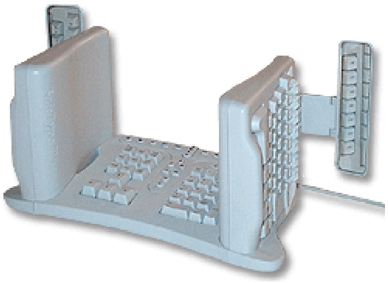

Figure 2.2 Vertical keyboard

Among the Fixed-Split Keyboards, Microsoft's Natural keyboard has done much to break the paradigm of what a keyboard should look like. Along with an earlier attempt by Apple's Adjustable keyboard, these mainstream names have largely legitimized the idea of alternative keyboards. According to the Washington Post (1996) Microsoft has accomplished a 61% share of the "ergonomic keyboard" market, with generic "home brands" making up an additional 24%.

2.2.1.2 Adjustable-Split Keyboards

Adjustable-Split Keyboards are able to change either their horizontal split or both the horizontal and vertical angling. The Comfort keyboard has been the higher-end of adjustable keyboards with the Goldtouch, Kinesis Maxim, and Pace keyboards being lower-cost alternatives. The most known split keyboard is Kinesis^{TM} keyboard, which is developed by Kinesis Corporation. The keyboard's design includes "a sculpted keying surface, separated alphanumeric keypads, thumb keypads, and closely placed function keys." The Kinesis keyboard puts keys in similar order with QWERTY keyboard, but arranges the keys for each finger in a vertical row to avoid the lateral hand motion when moving a finger from row to row. During the long time development, Kinesis keyboard adopt many ergonomic conceptions including contoured design, which will be introduced later. To some degree, Kinesis keyboard is a split keyboard as well as a contoured keyboard. Figure 2.3 shows some typical Kinesis keyboards in current market.

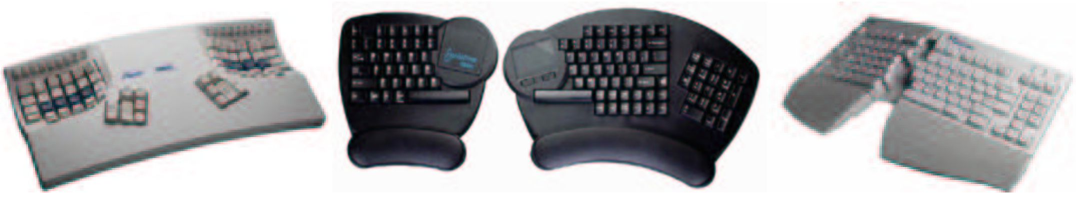

Figure 2.3 Some typical Kinesis keyboards (http://www.kinesis-ergo.com)

As shown in figure 2.3, Kinesis now has a two-piece keyboard with an integrated touchpad (left piece, right piece, or both). This design puts the keys in a way that corresponds to the shape of the hand. The keys for the middle finger are recessed more deeply, and the little finger keys are raised higher to shorten the finger motion in typing. In the conventional keyboards, the left thumb has nothing to do, and the right thumb just has one key- the spacebar. While in Kinesis keyboard, the right thumb covers six keys: space, Enter, Alt, Ctrl, Page Up, and Page Down. Space is the home position and Enter is reached by a slight extension of the thumb. The left thumb has its own Alt and Ctrl keys and also covers Delete, Home, and End. Backspace is the home position for the left thumb to correct errors without moving hand out of the home position.

One study conducted by Jahns, Litewka, Lunde, Farrand, and Hargreaves (1991), indicated that Kinesis muscle loads were substantially less than QWERTY muscles loads on muscles controlling hand deviation, extension, and pronation. In addition, participants indicated substantial preference for the Kinesis in areas of comfort, fatigue, and usability (Smith & Cronin, 1992).

2.2.2 Other Alternative Design

There are also other kinds alternative keyboards, one of which design places the letters in different places on the keyboard, more ergonomically set the keys in the curve most close the natural movement of operator's fingers which is named contoured keyboard. Usually it lessens the awkward postures associated with typing by changing the keyboard physical dimensions and layout (Honan et al.,1995). As for the current QWERTY keyboard, the distribution of letters for the English language are such that the left hand is active 60% and the less dominant fingers, such as the ring finger and the little finger, are recruited for many of the vowels. The most known contoured layout is Dvorak keyboard. In this keyboard layout it is more efficient for typing in the English language. (Jack Dennerlein, 2002)

2.2.2.1 Contoured Keyboards

Contoured Keyboards, also called sculpted keyboards, not only cut the standard keyboard into pieces and reassemble them but also place the keys in curves that closely match the natural movement of the fingers. By this way it reduces finger travel and also transfers some typing work from the weaker fingers, for example little finger, to multiple thumb keys. The most known contoured keyboard is Dvorak keyboard and its development, which were founded by Dvorak (1943) and improved by Kroemer (1972) and Hobday (1988) experienced couple of years.

A. Dvorak Keyboard

August Dvorak invented the Simplified Keyboard (as he called it) in 1932 as a result of exhaustive time and motion studies since he saw problems inherent in the QWERTY keyboard at his first sight. Those problems included not only limited type speed but also physical injuries, which are called symptoms Repetitive Stress Injury (RSI) today.

Dvorak Keyboard, as noticed previously, rearranged the alphabetic keys in a more ergonomic layout to distribute typing works more evenly among the fingers. As shown in figure 2.4, Dvorak's home row uses all five vowels and the five most common consonants: AOEUIDHTNS. According to the frequency, the vowels were placed on one side and consonants on the other. By strategic placement of the letters and punctuation, Dvorak typists are able to attain the same output more efficiently with reduced finger movement, thus reducing the strain on the hands, wrists, and arms. Due to its useful ergonomic features, it is accepted by the American National Standards Institute (ANSI). However, the retraining period for this keyboard was excessive according to Erdil and Dickerson's research (Erdil, M., Dickerson, O.B., 1997.). And the conventional QWERTY keyboard is so standard that it still in the charge of the market.

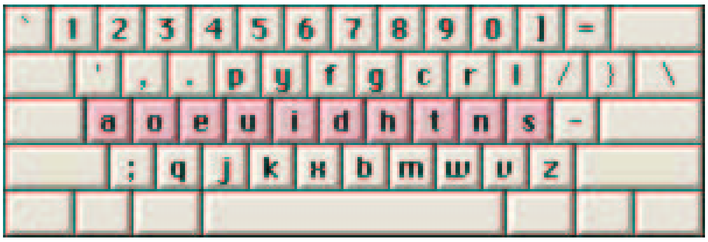

Figure 2.4The DvortyBoard keyboard layout

(http://www.mwbrooks.com/dvorak/layout.html)

After several decades' development, many new Dvorak keyboards are introduced today. Figure 2.5 shows some commercial models.

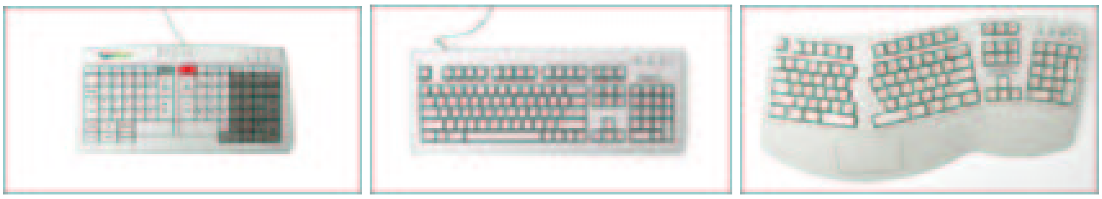

Figure 2.5 Dvorak/Qwerty Switchable Keyboards

(a) TypeMatrix 2020; (b) 2000 DQ; (c) 2001DQE

These advanced keyboards allow you to easily switch from the inefficient and exhausting Qwerty format to the efficient and comfortable Dvorak format by just touching the switch key. Even more, they are transparent to all applications and operating systems - even DOS.

B. Maltron Keyboard

In 1988, Hobby suggested a modified split key design based on Dvorak and Kroemer's work, known as the Maltron. The Maltron keyboard was also one kind of split keyboards, because it included a split key design to alleviate ulnar deviation as in the QWERTY layout. The numeric keypad was placed in the centre of keyboard, and more typing works are assigned to thumbs of both hands. As a contoured keyboard, it closely matches the finger length. A software conversion program was introduced in the keyboard design to make this design function with both the traditional QWERTY layout and an optimized layout. It associated the most commonly used keys such as vowels with the strongest and most appropriately positioned fingers.

2.2.2.2 Chording Keyboard

Chording Keyboards are another alternative to the standard keyboard. Chording keyboards are smaller and have fewer keys, typically one for each finger and possibly the thumbs. Instead of the usual sequential, one-at-a-time key presses, chording requires simultaneous key presses for each character typed, similar to playing a musical chord on a piano. Therefore chording keyboard requires far fewer keys than a conventional keyboard so that users can place the keyboard wherever it is convenient to avoid an unnatural keying posture (Cushman, W.H. & Rosenberg, D.J., 1991).

The typical chording keyboard is an alphanumeric input device, which is named the Alphanumeric Input Device for those with Carpal Tunnel Syndrome (AID-CTS) keyboard. It was developed specially to combat the problems of repetitive motion injury related to typing. The AID-CTS keyboard was designed aiming to eliminate finger movement, minimize wrist movement, and provide a more comfortable static posture for the hand. It uses a pair of devices each comprised of an inverted dome, which is coupled to a base.

In US, Kinesis is the better marketed and more popular version of these types of keyboards, especially when it comes to compatibility between many different computer platforms and providing for key and macro programmability. The Maltron keyboard was the pioneer in this style of keyboard and provides an optional, unique key layout. Its distribution seems to be more in Europe, but is also available in the US. The DataHand is a keying device that is the farthest from the traditional keyboard (short of chording devices) and is included in this category as it performs a similar function of limiting finger movement related to entering information into the computer. Among all these ergonomic keyboards, the MS Natural, Lexmark, and Kinesis keyboards have been the most popular of keyboards to first try out.

2.2.3 Ergonomic Keyboards under Development

Besides previous ergonomic keyboards, there are also various styles of keyboards under development. They are briefly introduced in following introduction.

E2 Solutions

The DataEgg, invented by Gary Friedman (Timothy Griffin, 2001), is currently being developed as a stand alone device. It is a round, one-handed, chording computer with a two-line LCD display. It can also serve as an alternative computer keyboard through a computers serial port (currently supporting the PC).

Ullman Keyboard

On the assumption that RSI in office work is mainly caused by to much static work and lack of dynamic work, the Ullman Keyboard (Timothy Griffin, 2001) was developed as an attempt to reduce the RSI problems, by minimizing the static muscular work needed to perform VDT work while maintaining the need for dynamic work. What it does is that just let the natural behaviour decide the design.

Keybowl – orbiTouch

The orbiTouch (Timothy Griffin, 2001) totally eliminates finger motion and wrist motion. A keystroke is created when operator slide the two domes into one of their eight respective positions. Hence sliding the domes to different positions inputs different letters and numbers. It is also the first ergonomically designed keyboard geared to all typists, especially those with Carpal Tunnel Syndrome (CTS) or other physical upper extremity disabilities.

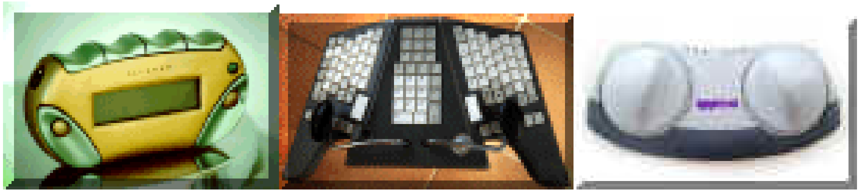

Figure 2.6 Keyboards under development

(a) E2 Solutions (b)Ullman Keyboard (c)Keybowl – orbiTouch

2.3 Voice Input Design

As discussed previously, a lot of research has been done to develop strike-key input method aimed to minimize WRSD development and improve work efficiency. Besides those keyboard ergonomic redesigns, voice input design is highlighted because of its hand free and high speed input characteristics. If voice input could be widely used, the ergonomic risk factors associated with keyboard would not exist at all. From this point of view, voice input would get rid of the risk factors radically. On the other hand, peoples, especially peoples with disabilities have huge hopes for operating their computers simply by speaking. This expectation became realistic with the rapidly development of speech technology. Automatic speech recognition (ASR) has already been used in a good many of applications, such as Web navigation, data entry, database access, browser and applet control, and remote control. Inspired with the great improvement, many pioneers made a great effort to use voice input instead of the conventional keyboard. To some degree, in no other area of technology has recent development been as dramatic as in the area of speech recognition. As the result, voice input systems become more and more numerous, and commercial advertising for these products becomes more and more pervasive.

There are several companies providing commercial voice recognition systems.

IBM (www.viavoice.ibm.com)

Dragon Systems (www.naturallyspeaking.com)

Lernout & Hauspie (www.ihs.com)

Phillips (www.vioce.be.phillips.com)

Among this wide range of products, two premier products in voice input technology are currently IBM's ViaVoice and Dragon Systems' Naturally Speaking series.

The ViaVoice family is an awarding-winning product line that takes advantage of the 40 years legacy of IBM speech research and development. The ViaVoice product family offers innovative features designed to make setup, dictation and voice navigation easier. The new ViaVoice version provides enhanced ease-of-use features for dictation and voice command of PC and Internet applications such as Email and Web navigation. Users can use voice to create, manage, and send email, chat on the Internet, command the browser, launch URLs and surf the Web. With ViaVoice, users can easily control the desktop and PC applications with voice by just saying the command name to activate menu options, lists and buttons. According to IBM's report, it has more than 300,000 vocabulary and backup dictionary words. All the ViaVoice products can be used for Microsoft Office XP, 2000 & 97, Outlook®, Internet Explorer, AOL and Netscape® Messenger®. Currently, the ViaVoice family has several versions to suit for different systems of both PC and Macintosh platforms.

Dragon Naturally Speaking is another ideal software for people to dictate text into standard applications so that users gain overall hands free from computer control. Its powerful scripting enables common tasks to be automated reducing workloads and dramatically increasing productivity. Similar to ViaVoice, Dragon Naturally Speaking enable users to dictate completely natural voices directly in to Microsoft Office and many other standard applications. The entry speech can be as high as 160 words per minute, and the accuracy can reach up to 95% according to official documents. With Dragon Natural Speaking, the user can control all aspects of computer usage through the voice, such as surf the Internet hands free. Also, the user can use a mobile recorder to create documents on the move, then connect the recorder to the computer and have Dragon transcribe the dictation. Besides the build in vocabulary, additional vocabularies are available to enhance the performance.

Both these two programs are based on speaker depended technique. In other words, they are trained to learn individual speakers' speech so that the programs can recognise individual speaker's voice and match the individual sounds to each word. The given routines in which users recite selected words and commands are helpful to get started. But the real training comes while dictating the real texts. Therefore users had better use the proper style all the time. Otherwise the program tends to misunderstand. On the other hand, since voice input is a delicate procedure, a slight change in ambient noise can affect the recognition accuracy. Unfortunately, such kinds of noise as a rasp in throat, a puff of air as exhale, or a minor background bang is unavoidable. This may lead an misunderstood word, i.e., "year" become "your," "either" become "air their," and so on. Any of actions of exhaling, cough or sneezing may lead to rather amusing results. Even though the programs claimed accuracy in the 90% range and 95% or better with practice, this data is only based on repetitious training and an ideal ambience without any noise. However, it is difficult for computer user to keep a completely quiet condition.

Furthermore, the correction process itself takes a bitter effort. Once an error occurs, the user has to speak a command, and then chose the right word from a menu with a numbered list of similar sounding words appeared on the screen. If the right word is there, the program replaces the original word. If the word isn't on the list, the users have to spell it one letter at time with the affiliated device. DragonDictate and Kurzweil Voice Pad use the military alphabet (Alpha, Bravo, Charlie, etc.) while as for IBM's Simply Speaking users have to type it in. It is obvious that users had better not completely liberated from the keyboard.

3 Bimodal Voice Recognition Based Input

3.1 Main Principle of Speech Recognition

In this thesis the recognition systems used for experiments were developed based on Hidden Markov Model (HMM) model. HMM approach is a well-known statistical method which is currently the most effective stochastic approach used to characterize the spectral properties of the frames of a pattern. After more than fifty years' research activity in speech recognition, HMM becomes one of the most successful approach to automatic speech recognition so far. Thus a brief review of the theory of HMM and its applications in the speech recognizer are introduced in the following part. S.J. Cox (1988), L. Rabiner (1989) and B. H. Juang (1993) introduced more detailed information in their articles and books.

3.1.1 Definitions

A hidden Markov model is a statistical model for an ordered sequence of variables, which can be well characterized as a parametric random process. It is assumed that the speech signal can be well characterized as a parametric random process and the parameters of the stochastic process can be determined in a precise, well-defined manner. Therefore, signal characteristics of a word will change to another basic speech unit as time increases, and it indicates a transition to another state with certain transition probability as defined by HMM. This observed sequence of observation vectors *O* can be denoted by

$$O = o(1),1(2),\ldots,o(T) \qquad (3.1)$$

where each observation $o(t)$ is an m-dimensional vector, extracted at time *t* with

$$o(t) = [o_1(t), o_2(t), ..., o_m(t)]^T \ . \qquad (3.2)$$

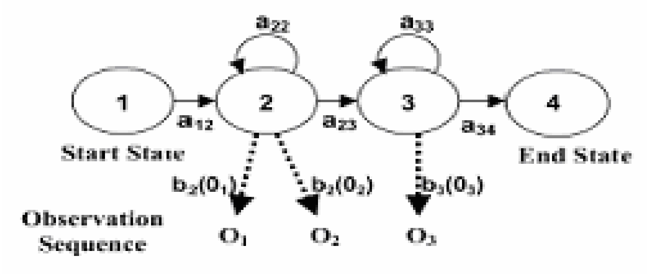

Figure 3.1 A typical left-right HMM (*aij* is the station transition probability from state *i* to state *j* ; *Ot* is the observation vector at time *t* and *bi*(*Ot*) is the probability that *Ot* is generated by state *i*).

3.1.1.1 Elements of a HMM

An HMM could be very complicated, but in general they can all be characterized by the following parameters:

a) *N,* the number of the states in the model. The states are hidden, however, each state within a process usually has some physical significance, like in the case of speech recognition, and each state could represent a basic speech unit. The states were denoted as $S = (s_1, s_2, ..., s_N)$ and the state at time *t* as q_t.

b) *M,* the number of the Gaussian mixture components per state, i.e., the discrete alphabet size. The individual symbols are denoted as $V = \{v_1, v_2, ..., v_M\}$.

c) *A*, the state transition probability distribution $A = \{a_{ij}\}$where

$$a_{ij} = P\left[q_{t+1} = s_j \middle| q_t = s_i\right], \quad 1 \le i, j \le N \tag{3.3}$$

the probability of being in state s_j at time $t+1$ given that we were in state s_i at time t and

$$\sum_{j=1}^{N} a_{ij} = 1, \qquad 1 \le i \le N . \tag{3.4}$$

There are many types of HMMs. For the special case such as ergodic model where all states can be reached by any other states, $a_{ij} \succ 0$ for all *i, j*.

d) B, for continuous HMMs, it is the matrix of observation probability distribution over all the state and all the observations. $B = \{b_j(k)\}$, where

$$b_j(k) = P\left[o_t = v_k \middle| q_t = s_j\right], \qquad \begin{matrix} 1 \le j \le N \\ 1 \le t \le T. \end{matrix} \tag{3.5}$$

$V = \{v_1, v_2, ..., v_M\}$, and

$$\sum_{t=1}^{T} b_j(t) = 1, \qquad 1 \le j \le N . \tag{3.6}$$

e) Π, the initial state distribution $\Pi = \{\pi_i\}$, in which

$$\pi_i = P[q_1 = s_i], \qquad 1 \le i \le N . \tag{3.7}$$

A complete specification of a HMM requires specification of two model parameters, *N* and *M*, specification of the observation symbols, and the specification of three sets of probability measures *A*, *B*, π_i. So an HMM can also be defined as a compact form $\lambda = \{A, B, \Pi\}$.

3.1.1.2 Three Problems for HMMs

In real applications, HMMs are used to solve three main problems. These problems are described as following:

Problem 1: Given the model $\lambda = \{A, B, \Pi\}$ and the observation sequence, how to efficiently compute $P(O|\lambda)$, the probability of occurrence of the observation sequence in the given model.

Problem 2: Given the model $\lambda = \{A, B, \Pi\}$ and the observation sequence, how to choose a optimal corresponding state sequence.

Problem 3: How to adjust the model parameters $\lambda = \{A, B, \Pi\}$ so that $P(O|\lambda)$ is maximized.

Problem 1 and problem 2 are analysis problems while problem 3 is a synthesis or model-training problem. To solve these three problems, some basic assumptions are being made in HMM.

a. *The output independence assumption:* The observation vectors are conditionally independent of the previously observed vectors.

b. *The stationary assumption:* It is assumed that state transition probabilities are independent of the actual time at which the transition takes place. It can be formulated mathematically as

$$P[q_{t1+1} = j | q_{t1} = i] = P[q_{t2+1} = j | q_{t2} = i] \tag{3.8}$$

for any t_1 and t_2.

3.1.2 Recognition

3.1.2.1 Baum-Welch Recognition

To find an optimized solution for problem 1, Baum-Welch algorithm or so-called Forward-Backward algorithm are adapted, which can efficiently calculate the likelihood over all the possible state sequences. The idea of the algorithm is that all possible sequences of the total likelihood must merge into one of the *N* states and the sum of the likelihood over all states at any time gives the total likelihood. In order to describe the Forward-Backward, such a recursive algorithm, two variables are introduced, the *forward probability* and *backward probability*.

The forward probability $\alpha_t(i)$ is defined as the joint probability of having generated the partial forward sequence up to the observation t and having arrived at state i:

$$\alpha_t(i) = P(o_1, o_2, ..., o_t, q_t = S_i | \lambda) \tag{3.9}$$

The forward probabilities can be calculated recursively using

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i)\alpha_{ij}\right] b_j(o_{t+1}), \qquad \begin{matrix} 1 \le t \le T-1 \\ 1 \le j \le N \end{matrix} \tag{3.10}$$

The initial condition is that:

$$\alpha_1(i) = \pi_i b_i(o_1) \qquad 1 \le i \le N. \tag{3.11}$$

The termination of the recursion is that:

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i)\,. \tag{3.12}$$

And the final probability can be given by

$$\alpha_T(i) = P(o_1, o_2, ..., o_T, q_T = s_i | \lambda) . \qquad (3.13)$$

Figure3.2 shows that how state S_i can be reached at time $t+1$ from the N possible states, S_i ($1 \le i \le N$) at time t. According to the definition of $\alpha_T(i)$ above, the Probability of observation $P(O|\lambda)$ can be achieved just as the sum of $\alpha_T(i)$'s.

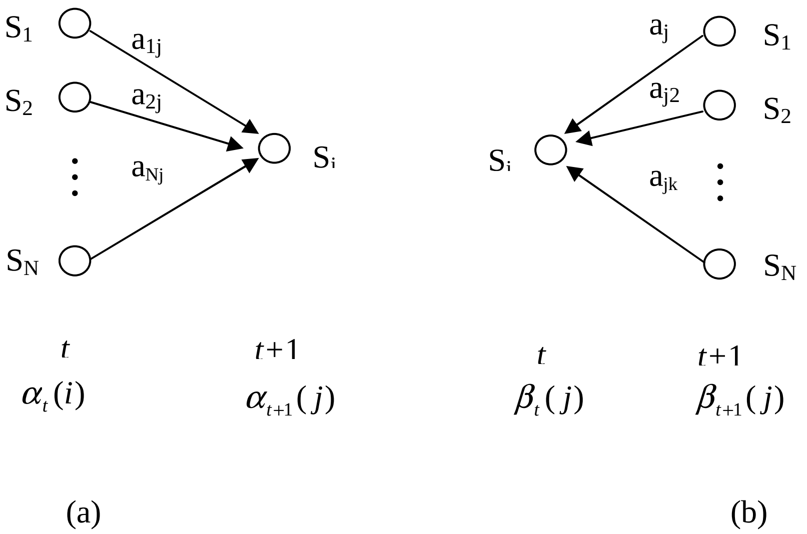

Figure 3.2 (a) Illustration of the sequence of operations required for the computation of the forward variable $\alpha_t(i)$ and (b) the computation of the backward variable $\beta_t(i)$ (L. Rabiner, 1989)

Figure 3.2 also shows the basic structure of backward variable $\beta_t(i)$, which is described as following:

$$\beta_t(i) = P(o_1, o_2, ..., o_T | q_t(i) = s_i, \lambda) \qquad (3.14)$$

Similar to the forward probabilities, the backward probability can be calculated recursively using

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j), \quad t = T-1, T-2, ..., 1, \quad 1 \le i \le N. \tag{3.15}$$

The initialization of backward variable is given by

$$\beta_T(i) = 1, \qquad 1 \le i \le N \ . \tag{3.16}$$

Defined

$$\beta_1(i) = P(o_2, o_3, ..., ...o_T | q_1 = s_i, \lambda), \tag{3.17}$$

The termination of the backward recursion is:

$$P(O|\lambda) = \sum_{i=1}^{N} \pi_i b_i(o_1) \beta_1(i) \ . \tag{3.18}$$

Therefore, the total likelihood is expressed by

$$P(O|\lambda) = \sum_{i-1}^{N} \alpha_t(i) \beta_t(i) \ . \tag{3.19}$$

3.1.2.2 Viterbi Recognition

As discussed in the previous section, the recognition of a word model using the Baum-Welch algorithm is based on the likelihood over all the possible state sequences. To find the best state sequence for a given observation sequence, i.e. to only consider the maximum likelihood state sequence, the Viterbi algorithm was introduced, as known as the solution of problem 2. The process of the Baum-Welch algorithm is to recognize the most likely word, while the Viterbi algorithm finds not only the most likely word but also the best state path (or state segmentation) of this word. In the

Viterbi algorithm, the likelihood is calculated using almost the same method as the forward probability calculation (Equation 3.10-3.13) except that the summation (3.11) is replaced by a maximum operation (3.21) in the backtracking step.

In order to find the single best state a variable *partial likelihood* for the Viterbi algorithm is defined as:

$$\delta_t(j) = \max_{q_1,q_2,\ldots,q_{t-1}} P\left[q_1, q_2, \ldots, q_t = j, o_1, \ldots o_2, \ldots, o_t \middle| \lambda\right] \tag{3.20}$$

where, $\delta_t(j)$ is the highest probability of the first *t* observations along a single state path, which ends in the state s_j at time *t*. Another variable to keep the trace of the state sequence is defined as $\psi_t(j)$. The recursive calculation of $\delta_t(j)$ and $\psi_t(j)$ is as follows:

$$\delta_t(j) = \max_{1\le i\le N}\left[\delta_{t-1}(i)a_{ij}\right]b_j(O_t), \qquad \begin{array}{c} 2 \le t \le T \\ 1 \le j \le N \end{array} \tag{3.21}$$

$$\Psi_t(j) = \arg\max_{1\le i\le N}\left[\delta_{t-1}(i)a_{ij}\right], \qquad \begin{array}{c} 2 \le t \le T \\ 1 \le j \le N. \end{array} \tag{3.22}$$

The initial conditions are:

$$\begin{array}{l} \delta_1(i) = \pi_i b_i(O_1), \quad 1 \le i \le N \\ \Psi_1(i) = 0. \end{array} \tag{3.23}$$

The termination of Viterbi algorithm is:

$$\begin{array}{l} P^{\bullet} = \max_{1\le i\le N}\left[\delta_T(i)\right] \\ q_T^{\bullet} = \arg\max_{1\le i\le N}\left[\delta_T(i)\right]. \end{array} \tag{3.24}$$

The backtracking of the state sequence is given as follows:

$$q_t^{\bullet} = \Psi_{t+1}\left(q_{t+1}^{\bullet}\right), \qquad t = T-1, T-2, \ldots, 1 \tag{3.25}$$

The recursion algorithm introduced above is the complete Viterbi algorithm. It can be seen that the Viterbi algorithm is similar in implementation to the forward calculation.

3.1.3 Parameters Re-estimation

In order to find a model that maximizes the probability of the given observation sequence, on the other word the solution of problem 3, Baum et al. (1970) defined an auxiliary function and proved the two propositions below:

Auxiliary function:

$$Q(\lambda, \overline{\lambda}) = \sum_{Q} P(O, Q|\lambda) \log P(O, Q|\overline{\lambda}), \tag{3.26}$$

where $\overline{\lambda}$ is the auxiliary variable that corresponds to λ.

Proposition 1. *If the value of $Q(\lambda, \overline{\lambda})$ increases, then the value of $P(O|\overline{\lambda})$ also increases,*

$$Q(\lambda, \overline{\lambda}) \geq Q(\lambda, \lambda) \rightarrow P(O|\overline{\lambda}) \geq P(O|\lambda). \tag{3.27}$$

Proposition 2. *λ is a critical point of $P(O|\lambda)$ if and only if it is a critical point of $Q(\lambda, \overline{\lambda})$ as a function of $\overline{\lambda}$,*

$$\frac{\partial P(O|\lambda)}{\partial \lambda_i} = \frac{\partial Q(\lambda, \overline{\lambda})}{\partial \overline{\lambda}_i}\Big|_{\overline{\lambda}=\lambda}, \quad 1 \leq i \leq D, \tag{3.28}$$

where D is the dimension of λ and $\lambda_i, 1 \leq i \leq D$, are individual elements of λ.

Furthermore, to describe the Baum-Welch reestimation algorithm, two variables are defined: *joint event* $\xi_t(i, j)$ and *state variable* $\gamma_t(i)$ (Figure3.2). $\xi_t(i, j)$ presents the probability of being in state s_i at time t and state s_j at time $t+1$ given the observation sequence O and model λ, i.e.

$$\xi_t(i, j) = P(q_t = S_i, q_{t+1} = S_j | O, \lambda) \tag{3.29}$$

With the definitions and theories of forward and backward variables in the previous section, $\xi_t(i, j)$ can be written in the form

$$\xi_t(i, j) = \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{P(O|\lambda)}. \tag{3.30}$$

State variable $\gamma_t(i)$ presents the probability of being in state s_i at time t given the observation sequence O and model λ:

$$\gamma_t(i) = P(q_t = S_i | O, \lambda) = \sum_{j=1}^{N} \xi_t(i, j). \tag{3.31}$$

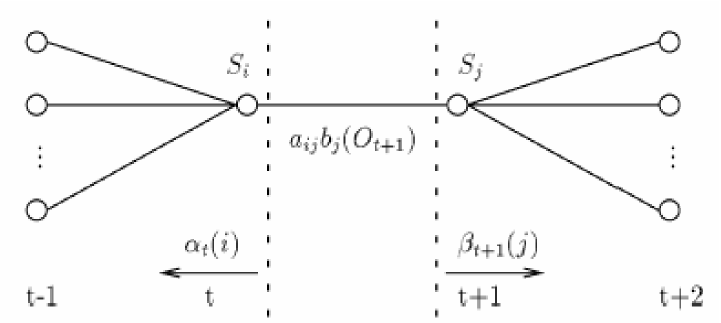

Figure 3.3Illustration of the sequence of operations required for the computational of the joint event that the system is in state s_i at time t and state s_j at time $t+1$ (L. Rabiner, 1989)

Using the above concepts, a method for reestimation of the parameters (Π, A and B)

of an HMM can been formed.

- $$\overline{\pi}_i = \text{expected } \textit{number of times in the state } S_i \textit{ at time } t=1 \\ = \gamma_1(i) \tag{3.32}$$

- $$\overline{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}, \quad 1 \le i \le N, \ 1 \le j \le N, \tag{3.33}$$

- $$\overline{b}_j(k) = \frac{\sum_{t=1, o_t=v_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}, \ 1 \le j \le N, \ 1 \le k \le M \ . \tag{3.34}$$

All the discussion above focus on the discrete symbols. For a continuous observation density, more parameters should be given the updating equations. We define density ϕ with mean vector μ and covariance matrix W, also define O as the vector being modeled and c_{jm} as the mixture coefficient for the *m*th mixture in state j.

Hence,

- $$\overline{c}_{jk} = \frac{\textit{probability of being in state j at time t with the kth mixture component}}{\textit{probability of being in state j}} \\ = \frac{\sum_{t=1}^{T} \gamma_t(j,k)}{\sum_{t=1}^{T} \sum_{k=1}^{M} \gamma_t(j,k)} \tag{3.35}$$

- $$\overline{\mu}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j,k) \cdot o_t}{\sum_{t=1}^{T} \gamma_t(j,k)} \tag{3.36}$$

- $$\overline{W}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j,k) \cdot (o_t - \mu_{jk}) \cdot (o_t - \mu_{jk})'}{\sum_{t=1}^{T} \gamma_t(j,k)} \tag{3.37}$$

- $$b_j(O) = \sum_{m=1}^{M} c_{jm} \phi\left[O, \mu_{jm}, W_{jm}\right] \tag{3.38}$$

3.2 Implementation of Bimodal Input

In the previous part, the basic principle of speech recognition was introduced. The whole recognition system is based on this theoretic foundation. However the whole implementation process is more than only finding solutions for a mathematic model, it can be described briefly as following procedures: Firstly, both visual and audio information are obtained. Then they are treated by different method to get the useful feature. Both of visual and audio feature go through the recognition engine and then results come out.

For audio signal, the general used acoustic features are Mel Frequency Cepstrum Coefficients (MFCC). This doesn't take a great effort, however, it is really challenge and difficult job to gain the visual features due to the high differences between and within speakers and the variability during speech production. Furthermore, the high variability of environment such as different illumination conditions cause further difficulties in image analysis. Large amounts of work have been carried out to get a robust and accurate visual feature analysis. These existing systems for extracting visual speech information from a sequence of images can be broadly classified according to which visual features they used for recognition and how to extract these features. They can be summarised as the following groups (S. Dupont and J. Luettin, 2000):

- Geometric-feature-based
- Model-based
- Image-based
- Visual-motion-based

In the *Geometric-feature-based* approach, it is assumed that certain measures such as the height or width of an opening mouth are important features. Most applications of this kind approach are semi-automatic methods or have painted the lips of the talker to facilitate feature extraction. Petajan (1985) developed one of the first audio video speech recognition systems. In this system, he used the geometric features such as the area, height and width of an opening mouth's image as the visual features. Goldschen (1994) developed a similar continuous lip-reading system. Further more, Chibelushi (1993) and J.S.D. Mason (1999) proved that the gross detail of a fine lip line provides much same information, therefore even very basic visual features can improve the recognition accuracy greatly when combined with acoustic signals. Obviously, it consumes less computational time and dimensional space than the image-based method which will be introduce later. In this method, the basic but essentially useful features are represented in a compact form. The disadvantage is the difficulty in automatic extraction of these features and the subjective choice of the features to consider.

The *model-based* approach (Kass 1988, Rao 1994, Coianiz et al., 1996, Robert Kaucic 1996, Luettin 1997, Basu 1998), on the other hand, is a model of the visible speech articulators. Usually the lip contours, is built and its configuration is described by a small set of parameters. In this approach, the deformable template model, and the active contours model, which includes the splines model (Barney Dalton and Andrew Blake 1996) and the snakes model (C. Bregler and S. M. Omohundro 1995), are used for lip tracking.

The advantage of the model based approach is that important features can be represented in a low-dimensional space and can often be made invariant to image transforms like translation, scaling, rotation and lighting. The disadvantage is that the particular model may not consider all relevant speech information. The main difficulty in the model-based approach is the definition of the model and the development of image search procedures that accurately find the correspondence between the model and the image.

In the *image-based* approach (B. P. Yuhas, et, at., 1990, C. Bregler and S. M. Omohundro, 1995, Peter L. Silsbee 1996, G. Potamianos, H. P. Graf, and E. Cosatto, 1998), the grey-level image containing speaker's mouth is either used directly or after some image transform as the feature vector. Whereas the *visual-motion-based* method (K. Mase and A. Pentland, 1991) assumes that visual motion during speech production contains relevant speech information. This approach can be considered as a sub group of the image based approach.

The advantage of these two methods is that no data about the visual speech is ignored. Obviously the disadvantage is that the feature based on images and visual-motions has a high dimensionality and there are a large number of feature vectors to be processed. Therefore, it is time-consuming. And the difficulty in collecting enough training data also leads to a high difficulty in speech modelling. Another disadvantage is that the classifier has to learn the task of finding the generalization for image variability and linguistic variability.

The follows described briefly how the system works: first, there are programs that allow the user to 'train' the computer to recognize his speaking at the first time using the system. Using the camera and microphone, a user reads the prompted words or sentence with a nature way. To get a satisfying result, it's better to repeat this once more or three times. Because speakers tend to slightly alter the way they speak. By repeating, computer tries to find a close match to the way the user generally speaking. Then the patterns are recorded. From now on, that particular sentence will act like a clicking on a shortcut key or using a macro and the designated action will be carried out. Of course, this bimodal voice recognition input system can be used as speaker independent, that is, different users can use the same one by skipping the training process. However, it will decrease the recognition accuracy, as the experiments showed in chapter 4.

4 Experimental Methods and Results

4.1 Acoustic-only Speech Recognition vs. Bimodal Speech Recognition

4.1.1 Objective

In the experiments, modelling of noisy environment and assessment of recognition accuracy are preformed. There are three main objectives as shown in following.

a. Compare the recognition accuracy of acoustic-only speech recognition and bimodal speech recognition for speaker-independent discrete words

b. Compare the recognition accuracy of acoustic-only speech recognition and bimodal speech recognition for speaker-independent continuous words.

c. Compare the recognition accuracy of acoustic-only speech recognition and bimodal speech recognition for speaker-dependent discrete words

4.1.2 Subjects

15 subjects are involved in the experiments. Among them, 4 are females and 11 are males. The average age of the subjects is 28.6. They are divided into three groups according to independent and dependent voice recognitions:

a. Group one: 5 subjects; Speaker-independent discrete words test.

b. Group two: 5 subjects; Speaker-independent continuous words test

c. Group three: 5 subjects; Speaker-dependent discrete words test

In order to simplify the problems, 10 words are used in the recognition procedures. They are zero, one, two, three, four, five, six, sever, eight and nine respectively.

The difference between continuous words and discrete words is that there is a longer pause between two adjoining words in discrete words test. The interval between two adjacent words is set as 3 seconds.

4.1.3 Experiment Hardware

Desktop with Pentium 4, 2.1G, 256M RAM, a Headset Microphone, Digital camera SONY DCR-PC100E

4.1.4 Experiment Software

Recognition software

In the experiments, two different recognition engines were used: Acoustic-only speech recognition and bimodal speech recognition. Both of them were developed by the speech group of National University of Singapore for research uses. In the bimodal speech recognition, Multi-dimensional MFCC vector are used as audio features. The width and the height of lip, and the angle on the corner of the outer lip contour are used as visual features.

Other assistant software

MATLAB 6.1, Microsoft Powerpoint 2002, Noise Editor (a software can separate audio part from AVI files and edit audio video files)

4.1.5 Experiment Mechanism

4.1.5.1 Continuous and Discrete words Generation

In order to generate continue and discrete words, MATLAB was used to program a small file, through which the order of the words can be generated randomly. In this experiment, 30 sets of data are generated and each set has 15 words. Among them, 20 sets of data are implied in speaker- independent test. The other 10 sets are for speaker-dependent continuous words test.

Microsoft PowerPoint is used to present data in front of subjects. In continuous words recognition test, all the 15 words were shown in one slide, so that the subject can pronounce all the words continuously. In discrete words recognition test, one slide only showed one word, and the interval between adjacent slides were set as 3 seconds so that the subjects could stop for a while before pronounce the next word.

4.1.5.2 Simulation of Noisy Environment

In order to quantify the noisy environment and control the signal to noise ratio (SNR), clean audio and video chip are recorded by the digital camera in a quiet class room, and then the clean speech audio signals were interrupted by white noise with different SNR which is calculated as the logarithm of the ratio between the average power of the speech signal and the white noise signal. All the noise simulation procedures were performed in the computer based on modification on audio part of AVI files through software NOICE EDITER. Finally, seven modified files were derived from each clean AVI file with SNR 0 dB, 5dB, 10dB, 15dB, 20 dB, 25dB and 30dB respectively. These modified files are the inputs for both acoustic-only speech recognition and bimodal speech recognition software.

4.1.5.3 Data Processing

The output of speech recognition software is text documents. Comparing the input content and the output content, the recognition errors are accounted and statistically analyzed. The recognition accuracy was calculated using the following equation.

$$R_{ac} = \frac{N_{error}}{N_{input}} \tag{4.1}$$

The average recognition accuracy and standard deviation (SD) were calculated for each group. Finally, comparing between acoustic-only and bimodal speech recognition were performed. Using T test method, statistical significance of the difference between two speech recognition methods was analyzed based on P value. In this thesis, critical P value is set as 0.01, i.e. two set data are significantly different when P is smaller than 0.01.

4.1.6 Experiments Procedure

Firstly, the audio and video file was recorded in a very quiet classroom. The subject is required to seat in front of the computer, wearing a headset microphone. (Caution: don't block the view from camera) The powerpoint shows the words which subjects should follow in a designed order. The camera was operated manually and only focused on the mouths of the subjects to record the shape and movement of the lips. Each subject needed to pronounce two sets of words. Therefore, ten AVI files are obtained from each test group.

Secondly, the audio parts are abstract from AVI files and are processed through software NIOCE EDITOR to add noise signals with SNR 0dB, 5dB, 10dB, 15dB, 20 dB, 25dB and 30dB respectively.

Finally, use the two speech recognition softwares to recognize the noicelized audio and AVI files.

4.1.7 Experiments Results

The experiments results are described as follows:

1. The comparison of recognition accuracy of acoustic-only speech recognition and bimodal speech recognition for speaker-independent discrete words was shown in Table 4.1.

Table 4.1 Speaker-independent recognition accuracy (%) for discrete words

SNR	0dB	5 dB	10 dB	15 dB	20 dB	25 dB	30dB
Acoustic RA %	60.00	83.33	92.00	94.00	94.67	95.33	95.33
Bimodal RA %	82.00	94.00	96.67	98.00	98.67	98.67	98.67
Improvement △%	36.67	12.80	5.07	4.26	4.23	3.50	3.50
P value	0.00002**	0.00419*	0.00477*	0.04056	0.02550	0.02609	0.04787

** illustrates P <0.001;* illustrates 0.001< P <0.01. RA = Recognition Accuracy

Table 4.1 demonstrated the significant improvement of accuracy as a result of the bimodal speech recognition, especially in noisy circumstance. The higher the noise is, the more the recognition accuracy is improved compared to the acoustic-only speech recognition. As shown in table 4.1, when noise level reaches 0dB, the accuracy of acoustic-only recognition drops badly to 60%, while the bimodal recognition accuracy keeps as high as 82%. P value corresponding to SNR 0dB is lower than 0.0001, which shows the recognition accuracy is significantly improved by bimodal speech recognition. When SNR are 5dB and 10dB, P value are smaller than 0.01. This also demonstrates there is significant difference between recognition accuracy of acoustic-only and bimodal speech recognitions. Therefore, combined visual speech information together with audio speech information, speech recognition becomes more robust to noise. This illustrated the significance of visual information in speech recognition.

Thus it is more practical for real use.

2. To know more about the system, acoustic-only speech recognition and bimodal speech recognition for speaker independent continuous words were examined. The results demonstrate again that the higher the noise is, the better the accuracy of bimodal recognition is compared to the acoustic only speech recognition. When noise level reaches 0dB and 5dB, P value are 0.0331 and 0.0434 respectively. 0.01<P<0.05 illustrate moderate evidence against the null hypothesis which is assumed that two set data are likely identical and there is no significant difference. Table 4.2 shows both the experiments results and P values.

Table 4.2 Speaker-independent recognition accuracy for continuous words

SNR	0dB	5 dB	10 dB	15 dB	20 dB	25 dB	30dB
Acoustic RA %	15.33	22.00	34.00	46.00	58.00	66.67	76.67
Bimodal RA %	20.00	27.33	38.00	50.00	60.00	70.67	77.33
Improvement △%	30.43	24.24	11.76	8.70	3.45	6.00	0.87
P value	0.0331	0.0434	0.1850	0.1566	0.1717	0.1483	0.4341

Although the highest accuracy is lower than that of discrete case, 77.33%, in continuous case, is still acceptable. If the speaker slow down the talking speed, and make more pauses between words, the accuracy will rise.

3. As pointed previously, the current commercial voice input software are speaker-dependent. To be more directly, the speaker-dependent recognition accuracy of both acoustic-only and bimodal recognition were examined. The results are shown in table 4.3. It illustrates again that the bimodal recognition is more robust to noisy speech signals (For SNR 0dB and 5dB, P<0.001).

Table 4.3 Speaker-dependent recognition accuracy (%) for discrete words

SNR	0dB	5 dB	10 dB	15 dB	20 dB	25 dB	30dB
Acoustic RA %	61.33	77.33	94.67	98.00	98.00	100.00	100.00
Bimodal RA %	82.67	94.00	98.67	99.33	99.33	100.00	100.00
Improvement △%	34.78	21.55	4.23	1.36	1.36	0.00	0.00
P Value	0.00047^{**}	0.00064^{**}	0.02550	0.17172	0.08393		

** illustrates P ＜0.001.

The experiments results for speaker dependent case are generally better than the ones for speaker independent case. This is understandable and reasonable. As introduced in chapter 3, the Hidden Markov Model is a statistical model. In the speaker dependent case, the system was trained first to get familiar with the speaker. Therefore, it is easier for the HMM based system to recognize the known speaker than to recognize a completely stranger, whose data are totally new for it.

Summarizing all experiments' results, Figure 4.1 shows the curves for recognition accuracy as a function of SNR.

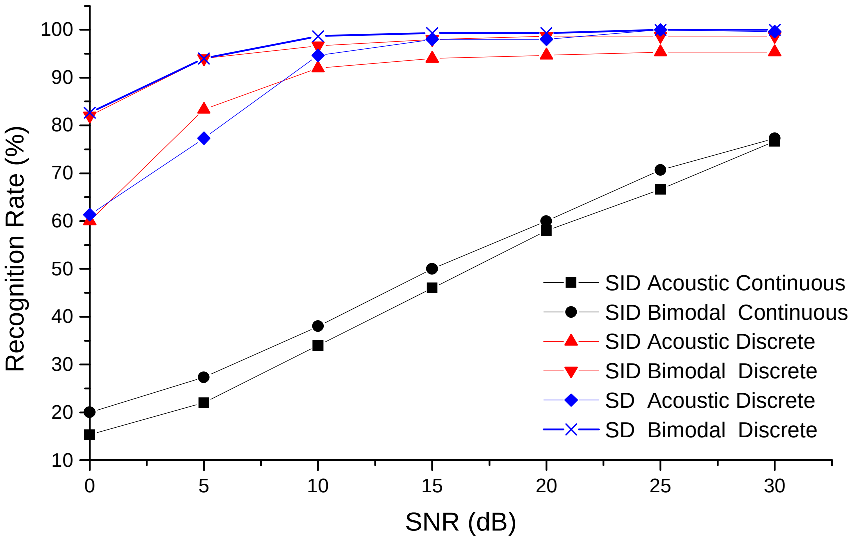

Figure 4.1 Recognition accuracy in different experiments

4.2 Voice Input Keyboard Design

Based on the data analysis in Chapter 4.1, bimodal speech recognition can significantly improve the perception of speech especially in noisy environment as shown in table 4.1 and 4.3. And also, it provides sufficient complementarity for audio signal. As one can image, different person speaks a word in different accent, but all speakers do a similar lip movement. Hence, integrating the bimodal speech recognizer to a keyboard would be very meaningful especially in noisy environment.

4.2.1 Design Methodology

Based on the engineering design science fundamental, the functional characteristics of keyboard are defined. Value analysis was used to weight different functions and rank different proposals to functions.

4.2.1.1 Identify the Function of Keyboards

Every product has several functions and in another way, functions provide rationale for the existence of products. The main function of the product is the most important rationale for the products (ARE053, 2003). Through the observations, brainstorming and the art of literature, the functions of the keyboard are identified as many as possible. Finally, the list of functions is given.

4.2.1.2 Weighting the Functions

15 functions of keyboard are summarized from the function list, which is obtained from step 4.2.1.1. In order to know which functions are important, comparison of these functions was carried out. Finally, the weight factor for every function is calculated from the ranking values. Via these weight factors, one can easily

distinguish which function play an important role in the new design procedure.

The weighting methods are based on the comparing between each function (D. Peterson, 2003). A ranking value 0, 1, 2 are given to each function according to its importance. For instance, four-function case are analysed in table 4.4. First, compare function 1 with the other functions (function 2, function3 and function 4). If function 1 is more important than function 2, the raking value for F1 to F2 is 2 as shown in table 4.4. Then, compare Function 1 with function 3. If the two functions are equal, the raking value for F1 to F3 is 1. Comparing Function 1 with function 4, if function 1 is less important than function 4, write down value 0. These procedures would be repeated until all the functions are compared with each other and all the ranking values R_{12} R_{13} R_{14} R_{23} R_{24} R_{34} are got as shown in table 4.4. Secondly, write down the correction term for each function. The correction term is introduced to balance each function. It can be obtained from equation 4.2.

$$CorrectionTerm\_for\_Function_i = 2\times(i-1) \qquad (4.2)$$

Thirdly, calculate the sum of vertical ranking values and time the sum by -1 for each function. This value is called negative ranking value. This value represents the total result of comparison between function (i) and function (1) to function (i-1).

Fourthly, calculate the sum of negative ranking value, positive ranking values and correction value for each function. For example, to add all the values in the first line, one can get weight number P_1 for function1; to add the values on second line, one can get weight number P_2 for function2…

Finally, weight factor Ki for function (i) is calculated according to equation 4.3

$$Ki = \frac{Pi}{\sum_{i=1}^{n} Pi} \qquad (4.3)$$

Table 4.4 Method to calculate weight factor for functions

	F1	F2	F3	F4	CT	Pi	Ki
F1		R_{12}=2	R_{13}=1	R_{14}=0	$2\times(1-1)=0$	3	0.25
	F2	$-(R_{12})$	R_{23}=2	R_{24}=1	$2\times(2-1)=2$	3	0.25
		F3	$-(R_{13}+R_{23})$	R_{34}=0	$2\times(3-1)=4$	1	0.083
			F4	$-(R_{34}+R_{24}+R_{14})$	$2\times(4-1)=6$	5	0.417
					Sum	$\sum Pi=12$	$\sum Ki=1$

4.2.1.3 Design Proposals

Based on the weighting results, the vital problems are defined in detail. Initially, several design proposals are generated. Then, two proposals, which adopt voice input technique, are selected and further modified to fulfil the new functions. And one ideal traditional ergonomic keyboard design was choose as reference. Between these two voice input proposals, one integrates acoustic-only speech recognition to keyboards, the other proposal integrates bimodal speech recognition to keyboards.

4.2.1.4 Proposal Ranking

As one can easily see, different proposal has different advantages and disadvantages. For example, a proposal maybe solve problem 1 very well, but doesn’t work at all in respect of problem 2. Hence, ranking proposals and get the best one is always the aim of developers.

In this step, the proposals are ranked using ranking method described as value analysis matrix. Firstly, one proposal is estimated to every function and ranked with point 0,1,2,3. Here, 0 means this proposal doesn't work at all for specified function and 1 is that it works but not well. 2 means that it works moderate well. Then, the final ranking value for this proposal is obtained by adding all point multiplied by function weight factor as described in equation 4.4.

$$Sum\_RankingValue = \sum_{i=1}^{n} K_i \times RP_i \qquad (4.4)$$

Where K_i is weighting factor for function (*i*).

RP$_i$ is the ranking point of the proposal for function(*i)*

n is the number of the functions.

The Previous procedure would be continued until the all proposals are ranked. Comparing the final ranking values, the advantages and disadvantages of different proposal are discussed and some recommendations are introduced to develop computer keyboard.

4.2.2 Design Results

4.2.2.1 Functions of Keyboard

1) Input letters, numbers and different symbols
2) Enable high speed input
3) Easy to strike the keys
4) Easy to remember the key positions
5) Suitable hand position angle
6) Movable
7) Cordless

8) Well-designed key size and shape

9) Comfortable tactile keys

10) Well-designed key layout to reduce wrist movements and repetitive movements (fatigue)

11) Enable hand free input

12) High reliability

13) Hot keys for internet, email , power management or own defined tasks

14) Quiet operation

15) Wrist rest

16) Palm rest

17) SP/2 or USB compatible

18) Saving space

19) Sturdy

20) Good looking

21) Easy plug and play

22) Tilting angle adjustable

23) Reduce errors

4.2.2.2 Weighting the Functions

Table 4.5 shows the weights of 15 functions, which are summarized from the previous step. The result shows that *'Input letters'* is the most important function, which have weight factor 0.127. The function *'Easy to remember the key position'* got second highest score 0.103. Then, *'Enable high speed input'*, *'Reduce fatigue'*, and *'hand free input'* have weight factors 0.099, 0.089 and 0.089 respectively, as the third and forth most important functions.

Table 4.5 Weight Factor of Functions

	High speed input	Key position	Cordless	Key size shape	Comfortable tactile	Reduce fatigue	Hand free	Input letters	Hot key	Wrist and Palm rest	Easy plug and play	Saving space	Sturdy	Good looking	Tilting angle	Correction Term	Summary	Weight Factor
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	I	Pi	Ki
A	-0	1	2	2	1	1	1	0	2	2	2	2	1	2	2	0	21	0.099
	B	-1	2	1	2	0	1	1	2	2	2	2	1	2	2	2	22	0.103
		C	-4	0	0	0	1	0	1	1	1	1	0	2	1	4	9	0.042
			D	-3	1	1	1	0	2	1	1	2	1	2	1	6	16	0.075
				E	-4	1	1	0	2	2	2	2	1	2	1	8	18	0.085
					F	-3	1	0	2	1	2	2	1	2	1	10	19	0.089
						G	-6	0	2	2	2	2	1	2	2	12	19	0.089
							H	-1	2	2	2	2	2	2	2	14	27	0.127
								I	-15	0	0	2	1	2	1	16	7	0.033
									J	-13	1	2	0	2	1	18	11	0.052
										K	-14	2	0	2	1	20	11	0.052
											L	-21	0	0	0	22	1	0.005
												M	-9	2	1	24	18	0.085
													N	-24	1	26	3	0.014
														O	-17	28	11	0.052
															Sum.		213	1

Compare A with B. If A is more important than B-write 2 point. If A is equal to B-write 1 point and if B is more important than A, write 0 point. Compare A with C, A with D etc

Pi is the sum of horizontal numbers

Weight Factor Ki= $\frac{Pi}{Sum\_of\_Pi}$

4.2.2.3 Design Proposals

Based on the weight factors of each function, Table 4.6 describes the functions in a descending order of the importance.

Table 4.6 Function Sorting

F. Code	Function	Weight Factor Ki
H	Input letters	0.127
B	Easy to remember the key positions	0.103
A	High speed input	0.099
G	Enable hand free input	0.089
F	Well-designed key layout to reduce repetitive wrist movements	0.089
E	Comfortable tactile keys	0.085
M	Sturdy	0.085
D	Well-designed key size and shape	0.075
J	Wrist and Palm rest	0.052
K	Easy plug and play rest	0.052
O	Tilting Angle adjustable	0.052
C	Cordless	0.042
I	Hot keys for internet, email and so on	0.033
N	Good looking	0.014
L	Saving space	0.005

According to the sorting result, the five most important functions are further considered as basic problems in the development of keyboard. Hence, the design problems are derived from these five basic functions.

1. Can the redesigned keyboards input letters as well as the traditional keyboards?

2. How to improve user’s input speed?

3. How to make user to be familiar with the key layout as soon as possible?

4. How to liberate our hand from repetitive typing?

5. How to avoid occupational health hazards

Aiming to solve the above problems, two solutions were proposed as shown in the following.

Proposal 1: Acoustic-only voice recognition computer keyboard

This design integrates a voice recognition chip and microphone into a keyboard. A switch was also designed on keyboard by which users can switch on and off voice recognition function. When the voice recognition function has been switched on, the computer would open the acoustic-only voice recognition software automatically. Once the software is ready, the cursor would change into a small mouth to notice user that the voice input function has started and please order the computer to do what they like. In this proposal, the traditional keyboard and mouse are still remained to act as a tool for modifying text, correct error, or be used in some special cases, such as library, where it is required to keep silence all the time.

As mentioned above, this proposal is based on combination of voice recognition and traditional keyboard. Even a traditional keyboard user can update their keyboard to acoustic-only voice recognition keyboard without throwing away anything or buying anything but the voice recognizer.

Proposal 2: Bimodal voice recognition keyboard

In this proposal, bimodal voice recognition is integrated to traditional keyboard. The difference between this and first proposal is that a camera is required here. Beside this, a similar switch key is added as in proposal 1, so that the user can choose any input method as they want. In this proposal, the keyboard and mouse also act as modification tools or alternative input method in special cases. Once error happens, the user can choose the most convenient way to correct according to the situation.

4.2.2.4 Ranking Proposals

The proposal 1 and proposal 2 are ranked 0,1,2,3 to the every function. 0 means that it doesn't work at all for that function and 1 is that it works but not well. 2 means that it works moderate well. 3 means that it works very well for that function. The detail ranking result is shown in table 4.7.

Because both of the proposals introduced voice recognition input method which can liberate our hand from endless typing, these solutions benefit for reducing hand movement and keep user far away from RTS and other keyboard related risks. Meanwhile, memorizing key layout is not as important as using traditional keyboards, because the input doesn't rely on finger striking keys any more.

The results in table 4.7 illustrate that both of the proposals got a higher score than the ideal traditional ergonomic keyboard. . The sum of point for proposal 2 is a little bit lower than proposal 1. This is mainly because proposal 2 requires a camera, and this increases the complexity for installing input devices. More detailed comparison and discussion were introduced in next chapter

Table 4.7 Proposal Rating

		FUNCTION															
		High speed input	East to memorize key positions	Cordless	Key size shape	Comfortable tactile key	Reduce fatigue	Hand free	Input letters	Hot key	Wrist and palm rest	Easy plug and play	Saving space	Sturdy	Good looking	Tilting angle	Sum of point
Weight Factor		0.099	0.103	0.042	0.075	0.085	0.089	0.089	0.127	0.033	0.052	0.052	0.005	0.085	0.014	0.052	
Prop 1	RP_i	3	2	0	2	3	3	3	3	3	3	3	0	3	3	3	2.687
	$K_i \times RP_i$	0.297	0.206	0	0.15	0.255	0.267	0.267	0.381	0.099	0.156	0.156	0	0.255	0.042	0.156	
Prop 2	RP_i	3	2	0	2	3	3	3	3	3	3	2	0	3	3	3	2.635
	$K_i \times RP_i$	0.297	0.206	0	0.15	0.255	0.267	0.267	0.381	0.099	0.156	0.104	0	0.255	0.042	0.156	
Traditional Keyboard	RP_i	1	1	3	2	3	0	0	3	3	3	3	1	3	3	3	1.983
	$K_i \times RP_i$	0.099	0.103	0.126	0.15	0.255	0	0	0.381	0.099	0.156	0.156	0.005	0.255	0.042	0.156	

5 General Discussion

5.1 Compare Voice Input Keyboard with Traditional Keyboard

Although different kinds of ergonomic keyboards were introduced, and more new designs are in developing, none of these devices have been broadly adopted. First reason is the learning curve associated with adopting a new keyboard is critical. Secondly, musculoskeletal disorder still exists even in the new ergonomic keyboard. For example, keyboard layouts were optimized in an effort to take advantage the stronger fingers. However, this is only suitable in the individual case. In the highly repetitive typing task, these fingers have to endure times over fatigue and finally suffer the injury due to the gradual deterioration. In this thesis two voice recognition based computer input methods were proposed. Both two methods can avoid repetitive movement of fingers as much as possible. And also it can give users freedom to move their hand away from the keyboard. From this point of view, voice input method is more efficient than traditional ergonomic keyboard to keep computer users far away from repetitive motion injures.

Another significant advantage of voice input methods is their huge benefit for disable persons. Users with physical disabilities who can only control the computer with one or two specific movements can speak to their computers and have them type by themselves. Even for deaf people, they could use the voice input function to take note in classes and let the computer tell them what teacher is talking.

Third advantage of these two proposals is high-speed input. Please imagine, why in our daily life, we prefer talk with each other to writing letters in most of the time. The main reason is that voice communication is the easiest, fastest and most direct way in most of time comparing to other communication methods. Similar is communication with computer. Voice input method is several times faster than traditional keyboard input. No doubt, it is the natural, easy and fast human-computer communication way, which mirrors our normal social interaction.

The experiments results also showed that voice input keyboards have significant advantage over the traditional keyboards. The ranking value for two voice input keyboard is about 34% higher than that for traditional ergonomic keyboard.

With regard to typing accuracy, voice input has to lose in front of traditional keyboard input. Some problems inherently exist and are hard to overcome because of homophones in our language. In this case, and interactive window has to pop up and ask user to choose the right one.

The other disadvantage of voice input is the recognition software may occupy a lot of computer resource comparing with traditional keyboard input. These may cause system resource exhausted. Only making the computer processor more powerful and memory larger can solve this problem.

Another disadvantage of voice input is that it is easy to disturb the other person, such as co-worker etc. In multiple office (more than 2 person share one office room), in order not to disturb the other, the users had better use the keyboard input.

Although voice input keyboard may avoid a lot of work related health hazards, no one is certain that it will not lead to any new potential injury such as pharyngitis etc.

Because the existence of so many shortcomings of voice input, the traditional keyboards still play a great role in computer stage. And the redesigned the proposals have to remain keyboard and mouse as an alternative and assistant input method as they are still in the early stage.

5.2 Compare Bimodal Voice Input Keyboard with Acoustic-only Voice Input Keyboard

As mentioned in Chapter 4, bimodal speech recognition can significantly improve the perception of speech, especially in noisy environment. Results from experiment 4.1 shows that as the noise level increases, the audio-only recognition decreases to unacceptable low levels. Hence, the visual speech features provide the complementary and necessary information useful to improve the recognition accuracy in bimodal case. However the bimodal speech recognition also has its own disadvantage. Firstly, in high SNR level, the importance of visual information becomes less significant. As the experiment results showed, at the noise level of 30dB (SNR), the acoustic only recognition accuracy begins to get the same accuracy as the bimodal recognition can. Secondly, the visual speech features are more complicated and not as robust to variance of the speaker as the acoustic features were. As the bimodal recognition experiment results for speaker independent task showed, the recognition accuracy is dissatisfied as only 77.3% at noise level 30dB. To enhance the robustness for speaker variance in bimodal speech recognition, it requires some further research.

Another point, which we should pay attention, is the results in ranking proposal step. Because tradition keyboards do not get influence from environmental noise, the noise factor is not include in keyboard functions. Therefore, the proposal 2 did not show its advantage in ranking step. Instead, the ranking points for proposal 2 is lower than proposal 1 because of system complexity.

Comparing with acoustic-only voice input, bimodal voice input needs larger memory and CUP resources. Beside this, a camera is required to get the visual information.

Currently, many laptops such as Sony have their own build-in camera. However, most desktops don’t have it. Requirement of camera would increase the cost for bimodal voice input keyboard.

6 Conclusions and Recommendations

In this thesis, a bimodal approach of voice recognition based computer input method was introduced. The voice recognition experiments shows that combined visual speech information together with audio speech information, bimodal speech recognition are more robust to noise. The significant difference ($P<0.001$) between acoustic and bimodal voice recognition is found when the SNR is 0dB and 5dB. Therefore, the anti-noise ability of bimodal voice recognition system is higher than acoustic-only voice recognition system.

Based on two kind of voice recognition system, two proposals are proposed. Then, the two proposals are analyzed and compared with traditional ergonomic keyboard design using a systematic product design method. The results shows both of the two redesigned voice input keyboards have significant advantages over the traditional ergonomic keyboard. The overall ranking values for voice input keyboard are 32.7% and 38.6% higher than that of traditional ergonomic keyboard. Therefore, voice input keyboard are highly recommended for the persons who are bothered by repetitive strain or stress injuries (RSI), work-related upper extremity disorders (WRUED) and disabilities.

However, as these methods are still in their early stage, there are some disadvantages too. More research is required to make them more practical.

7 Reference

B. P. Yuhas, M. H. Goldstein, T. J. Sejnowski, and R. E. Jenkins (1990), Neural network models of sensory integration for improved vowel recognition, Proc. IEEE, vol. 78, pp. 1658–1668

B. Yuhas, J. Goldstein, T. Sejnowski, and R. Jenkins(1990), ‘Neural network models of sensory integration for improved vowel recognition, Proceedings of the IEEE, vol. 78, no. 10, pp. 1658–68

Bailey, R.W. (1982). Human Performance Engineering: A guide for system designers. Englewood Cliffs, NJ: Prentice-Hall.

Bergqvist U, Wolgast E, Nilsson B, Voss M (1995a) Musculoskeletal disorders among visual display terminal workers: individual, ergonomic and work organizational factors. Ergonomics, 38(4): 763-776

Bergqvist U, Wolgast E, Nilsson B, Voss M (1995b) The influence of VDT work on musculoskeletal disorders. Ergonomics, 38(4): 754-762

Bureau of Labor Statistics (BLS) (2001): Reports on Survey of Occupational Injuries and Illnesses in 1977- 2000. Washington. DC: Bureau of Labor Statistics, US Dept. of Labor.

C. Bregler and S. M. Omohundro.(1995), Nonlinear manifold learning for visual speech recognition . In IEEE International Conference on Computer Vision, pages 494-499, IEEE, Piscataway, NJ, USA

C. Chibelushi, J. Mason, and F. Deravi.(1993), Integration of acoustic and visual speech for speaker recognition . In *Proc. EUROSPEECH*, pp. 157-160

Carter, J.B., Banister, E.W.(1994), Musculoskeletal problems in VDT work: a review. Ergonomics 37 (10), 1623-1648.

Chien-Yi Lu, J.(1997), Using Muscle Twitch to Measure Muscle Fatigue in Forearm Extensor Muscles During Typing, Masters Thesis, University of California, Berkeley

Cushman, W.H. & Rosenberg, D.J., (1991). Advances in Human Factors/Ergonomics 14: Human Factors in Product Design. New York: Elsevier. pp. 179-193.

D. Petterson (2003), Handout of Industrial Design Methodology, ARE053 master program of Ergonomics

Dennerlein JT, Yang MC. (2001) Haptic force-feedback devices for the office computer: Performance and musculoskeletal loading issues. Human Factors, 43(2): 278-86.

E. D. Petajan(1985), Automatic lipreading to enhance speech recognition, in Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 40–47

Erdil, M., Dickerson, O.B.(1997), Cumulative Trauma Disorders. Prevention, Evaluation and Treatment. Van Nostrand Reinhold, New York

Faucett, J., Rempel, D. (1994) VDT-Related Musculoskeletal Symptoms: Interactions Between Work Posture And Psychosocial Work Factors. American J. Industrial Medicine, 26(5):597-612.

Fernström, E., Ericson, M.O., Malker, H.(1994), Electromyographic activity during typewriter and keyboard use. Ergonomics 37 (3), 477-484.

Feuerstein M., Armstrong, T. J., and Hickey, P.(1994), Keyboard Force, Fatigue and

Pain in Symptomatic and Asymptomatic Word processors, 12th Congress of the International Ergonomics Association, Toronto, Canada, August 15-19

Feuerstein, M., Armstrong, T. J., Hickey, P., and Lincoln, A.(1997), Computer Keyboard Force and Upper Extremity Symptoms, Journal of Occupational and Environmental Medicine, Volume 39, Number 12

Franzblau A, Flashner D, Albers JW, Blitz S, Werner R, Amstrong T (1993) Medical screening of office workers for upper extremity cumulative trauma disorders, Archives of Environmental Health 48: 164-170.

G. Potamianos, H. P. Graf, and E. Cosatto (1998), An image transform approach for HMM based automatic lipreading, in Proc. IEEE Int. Conf Image Processing, pp. 173–177,

Gerard M.J. Gerard, S.K. Jones, L.A. Smith, R.E. Thomas and T. Wang (1994) , An ergonomic evaluation of the kinesis ergonomic computer keyboard. Ergonomics 37 10, pp. 1661–1668.

Gerr F, Marcus M, Ensor C, Kleinbaum D, Cohen S, Edwards A, Gentry E, Ortiz DJ, Monteilh C. (2002) A prospective study of computer users: I. Study design and incidence of musculoskeletal symptoms and disorders. Am J Ind Med. 41:221-235.

Goldschen, A.J., Garcia, O.N. & Petajan, E. (1994), Continuous optical automatic speech recognition by lipreading. 28th Annual Asilomar Conference on Signals, Systems, and Computers, Oct 31-Nov 2, 1994, Pacific Grove, CA.

Hedge, A., Powers, J.R.(1995), Wrist postures while keyboarding: effects of a negative slope keyboard system and full motion arm supports. Ergonomics 38 (3), 508-517.

Honan MM, Serina E, Tal R, Rempel D. (1995) Wrist Postures While Typing On A Standard And Split Keyboard. Proc. of the Human Factors and Ergonomics Society 39th Annual Meeting, San Diego, CA

http://www.aopd.com/vdt.html

http://www.kinesis-ergo.com/

http://www.library.wisc.edu/etext/WIReader/Images/WER0841.html

http://www.mwbrooks.com/dvorak/layout.html

Hunting, W., Laubli, Th. and Grandjean, E(1983). Constrained postures of VDU operators. Ergonomic Aspects of Visual Display Terminals, Taylor&Francis, London.

J. Luettin and N. A. Thacker(1997), Speechreading using probabilistic models, Comput. Vis. Image Understand., vol. 65, no. 2, pp. 163–178

J. Movellan(1995), ‘Visual speech recognition with stochastic networks, Proceedings of NIPS94 - Neural Information Processing Systems: Natural and Synthetic, pp. 851–8

Jack Dennerlein(2002), Musculoskeletal Disorders and the Computer Workstation: Research Supporting Ergonomic Interventions, December

K. Mase and A. Pentland (1991), Automatic lipreading by optical flow analysis, Syst. Comput. Jpn., vol. 22, no. 6,.

Kroemer, K.H.E. (1972). Human engineering the keyboard. Human Factors, 14, 51-63.

L. E. Baum, T. Petrie, G. Soules, and N. Weiss (1970), “A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains”, Annals of Math. Statistics, vol. 41, no. 1, pp. 164-171

L. Rabiner and B. H. Juang(1993), Fundamentals of speech recognition, Prentice Hall Signal Processing Series

L. Rabiner(1989), A tutorial on hidden markov models and selected applications in speech recognition, Proceedings of the IEEE, vol. 77, no. 2, pp. 257–86

Lueder, R.K. (1985) How to use computers without using keyboards. Office Ergonomics Review, 2(1), 22-24.

M. Kass. al, A.Witkin, and D. Terzopoulos. Snakes(1988), Active Contour Models. International Journal of Computer Vision, 1(4): 321–331

Martin BJ, Armstrong TJ, Foulke JA, Natarajan S, Klinenberg E., Serina E., Rempel D.(1996), Keyboard Reaction Force and Finger Flexor Electromyograms During Computer Keyboard Work. Human Factors , 38:654-664.

Martin, B. J., & Rempel, D. M.(1996), Muscle Activity During Computer Input Device Use, International Conference on Occupational Disorders of the Upper Extremities, Ann Arbor, Michigan, October 24-25, 1996.

Nakaseko, M., Grandjean, E., Hunting, W., & Gierer, R. (1985). Studies on ergonomically designed alphanumeric keyboards. Human Factors, 27, 175-187.

National Institute for Occupational Safety and Health, HETA 89-250.

National Institute for Occupational Safety and Health, HETA 90-013.

Noyes, J. (1983) The QWERTY keyboard: A review. International Journal of Man-

Machine Studies, 18, 265-281.

Pascarelli, E.F., Kella, J.J.(1993), Soft-tissue injuries related to use of the computer keyboard: A clinical study of 53 severely injured persons. Journal of Occupational and Environmental Medicine 35 (5), 522-532.

Peter L. Silsbee, Alan C. Bovik (Sep, 1996), ‘Computer lipreading for improved accuracy in automatic speech recognition , IEEE Transactions on Speech and Audio Processing, vol. 4, No.5, pp. 337–351

R. A. Rao, Mersereau, R.M.(1994), Lip modeling for visual speech reconition , Signals, Systems and Computers, 1994. 1994 Conference Record of the Twenty-Eighth Asilomar Conference on, Page(s): 587 -590 vol.1

Robert Kaucic, Barney Dalton, and Andrew Blake(1996), Real-time lip tracking for Audio-Visual speech recognition applications . In Proceeding of the European conference on Computer Vision, volume 2,pages 376-386, Cambridge

S. Basu, N. Oliver, and A. Pentland(1998), 3D modeling and tracking of human lip motion, in Proc. IEEE Int. Conf. Computer Vision,

S. Dupont and J. Luettin (2000), “Audio-visual speech modeling for continuous speech recognition,” IEEE Transactions on Multimedia, vol. 2, No. 3, pp. 141–51

S.J.Cox(Apr 1988), Hidden markov models for automatic speech recognition: theory and application, British Telecom Technical Journal, vol. 6, no. 2, pp. 105–115

Sauter SL, Schleifer LM, Knutson SJ. (1991). Work posture, workstation design, and musculoskeletal discomfort in a VDT data entry task. Hum Factors 33(2): 151-67.

Smith, W. and Cronin, D. (1992). Ergonomic test of the Kinesis keyboard. Global Ergonomic Technologies, Inc. Independent Study.

T. Coianiz, L. Torresani, and B. Capril (1996), 2D deformable models for visual speech analysis, in Speechreading by Humans and Machines: Models, Systems and Applications, D. G. Stork and M. E. Hennecke, Eds. Berlin, Germany: Springer-Verlag, vol. 150 of NATO ASI Series, Series F: Computer and Systems Sciences, pp. 391–398.

Timothy Griffin, 2001, http://tim.griffins.ca/gallery/keyboard

Tittiranonda et al. (1999) Tittiranonda P, Rempel D, Armstrong T, Burastero S. Effect of four computer keyboards in computer users with upper extremity musculoskeletal disorders. Am J Ind Med. 35(6):647-61.

Tittiranonda P, Rempel D, Armstrong T, Burastero S.(1997) Effect of four computer keyboards in computer users with upper extremity musculoskeletal disorders. Am J Ind Med. 35(6):647-61.

U.S. Occupational Safety and Health Administration (1997) Working Safely with Video Display Terminals (OSHA Publication 3092): http://www.osha.gov/

William Lehr (June, 1998), Computer use and productivity growth in US federal government agencies, 1987-92, JINDE, Vol.46, No.2:257-279

Wright, K.S. (August, 1996). Characteristics of Alternative Keyboard Acquisition, Setup, Use, and Benefits: A Survey Study. Masters Thesis, Human Factors & Ergonomics Program at San Jose State University. San Jose, CA: SJSU.

Wright, K.S., Andre, A.D. (1996). Alternative keyboard characteristics: A survey study. Conference Proceedings: ErgoCon'96 (pp 148-157). San Jose, CA: Silicon

Valley Ergonomics Institute.

Wright, K.S., Andre, A.D. (Oct/Nov 1997). Alternative keyboard purchasing decisions. Workplace Ergonomics (pp 28-32). Mount Morris, IL: Stevens Publishing Corp.

Appendix A

Table A **Group 1**

5 subjects; Speaker-independent discrete words test.

Acoustic	Number of words recognized correctly						
	0dB	5dB	10dB	15dB	20dB	25dB	30dB
Subject 1	9	13	14	15	15	15	15
	10	11	14	15	14	15	14
Subject 2	8	10	14	13	13	14	14
	8	14	14	14	14	14	14
Subject 3	8	13	14	15	15	15	15
	10	14	13	15	15	14	15
Subject 4	11	13	15	14	14	14	14
	10	12	12	13	14	14	14
Subject 5	9	12	14	13	14	14	14
	7	13	14	14	14	14	14
Average	9	12.5	13.8	14.1	14.2	14.3	14.3
RA	0.6	0.8333	0.92	0.94	0.94667	0.95333	0.95333

Bimodal	Number of words recognized correctly						
	0dB	5dB	10dB	15dB	20dB	25dB	30dB
Subject 1	12	15	15	15	15	15	15
	13	15	15	15	15	15	15
Subject 2	14	14	15	15	15	15	15
	13	14	15	15	15	15	15
Subject 3	11	15	15	15	15	14	14
	11	15	14	14	14	14	14
Subject 4	13	15	15	14	14	15	15
	13	12	13	14	15	15	15
Subject 5	12	13	15	15	15	15	15
	11	13	13	15	15	15	15
Average	12.3	14.1	14.5	14.7	14.8	14.8	14.8
RA	0.82	0.94	0.96667	0.98	0.98667	0.98667	0.98667
Improvement	0.36667	0.128	0.05073	0.04255	0.04226	0.03497	0.03497
P Value	2.1439E-05	0.00419	0.004767	0.040563	0.025502	0.026089	0.047867

Table B **Group 2**

5 subjects; Speaker-independent continuous words test

	Number of words recognized correctly						
Acoustic	0dB	5dB	10dB	15dB	20dB	25dB	30dB
Subject 1	3	3	5	8	9	10	12
	2	3	7	7	10	11	13
Subject 2	3	3	7	9	8	9	11
	2	4	6	6	9	13	14
Subject 3	3	5	5	8	9	10	11
	1	3	5	6	8	10	12
Subject 4	2	3	4	7	9	10	12
	3	4	6	6	10	11	12
Subject 5	3	2	3	7	10	10	10
	1	3	3	5	5	6	8
Average	2.3	3.3	5.1	6.9	8.7	10	11.5
RA	0.15333	0.22	0.34	0.46	0.58	0.66667	0.76667

Bimodal	Number of words recognized correctly						
	0dB	5dB	10dB	15dB	20dB	25dB	30dB
Subject 1	4	4	5	7	9	12	13
	2	3	6	7	10	11	12
Subject 2	3	4	6	6	7	10	13
	3	4	6	8	9	10	10
Subject 3	3	3	3	9	10	11	12
	4	5	7	9	10	12	13
Subject 4	2	5	6	7	10	10	13
	3	6	6	8	10	10	10
Subject 5	3	4	8	9	9	11	11
	3	3	4	5	6	9	9
Average	3	4.1	5.7	7.5	9	10.6	11.6
RA	0.2	0.27333	0.38	0.5	0.6	0.70667	0.77333
Improvement	0.304348	0.242424	0.117647	0.086957	0.034483	0.06	0.008696
P Value	0.03311	0.043421	0.185042	0.156642	0.171718	0.148333	0.434132

Table C **Group 3**

5 subjects; Speaker-dependent discrete words test

	Number of words recognized correctly						
Acoustic	0dB	5dB	10dB	15dB	20dB	25dB	30dB
Subject 1	11	12	15	15	15	15	15
	9	13	15	15	14	15	15
Subject 2	7	11	13	14	15	15	15
	9	10	14	14	14	15	15
Subject 3	8	12	14	15	15	15	15
	11	11	15	15	15	15	15
Subject 4	7	10	14	14	14	15	15
	9	12	14	15	15	15	15
Subject 5	10	12	14	15	15	15	15
	11	13	14	15	15	15	15
Average	9.2	11.6	14.2	14.7	14.7	15	15
RA	0.613333	0.773333	0.946667	0.98	0.98	1	1

Bimodal	Number of words recognized correctly						
	0dB	5dB	10dB	15dB	20dB	25dB	30dB
Subject 1	12	15	15	15	15	15	15
	13	15	15	15	15	15	15
Subject 2	13	14	15	15	15	15	15
	12	14	15	15	15	15	15
Subject 3	13	15	15	15	15	15	15
	12	15	15	15	15	15	15
Subject 4	13	15	15	15	14	15	15
	12	12	13	14	15	15	15
Subject 5	13	13	15	15	15	15	15
	11	13	15	15	15	15	15
Average	12.4	14.1	14.8	14.9	14.9	15	15
RA	0.826667	0.94	0.986667	0.993333	0.993333	1	1
Improvement	0.347826	0.215517	0.042254	0.013605	0.013605	0	0
P Value	0.000471	0.000639	0.025502	0.171718	0.083925		